

## A CAT with caveats: is the Consensual Assessment Technique a reliable measure of graphic design creativity?

Karl K. Jeffries

To cite this article: Karl K. Jeffries (2017) A CAT with caveats: is the Consensual Assessment Technique a reliable measure of graphic design creativity?, International Journal of Design Creativity and Innovation, 5:1-2, 16-28, DOI: [10.1080/21650349.2015.1084893](https://doi.org/10.1080/21650349.2015.1084893)

To link to this article: <https://doi.org/10.1080/21650349.2015.1084893>



© 2015 The Author(s). Published by Informa UK Limited, trading as Taylor & Francis Group



Published online: 14 Sep 2015.



Submit your article to this journal [↗](#)



Article views: 1331



View related articles [↗](#)



View Crossmark data [↗](#)



Citing articles: 1 View citing articles [↗](#)



# A CAT with caveats: is the Consensual Assessment Technique a reliable measure of graphic design creativity?

Karl K. Jeffries

School of Art, Design & Performance, University of Central Lancashire, Preston, UK

## ABSTRACT

The Consensual Assessment Technique (CAT) is considered one of the gold standards for creativity assessment, and graphic design, arguably, is the most ubiquitous domain within the creative industries. For the first time, this study tests two tasks to measure graphic design creativity, and by extension, the reliability of the CAT as a measure of graphic design creativity. Initial research suggested the level of consensus amongst judges (often referred to as inter-rater reliability) was too low to be reliable, and may be unduly influenced by a judge's preference for technical execution. In this study, 16 professional graphic designers were randomly assigned instructions to discount technical execution from creativity ratings, or given instruction that gave no stipulation, for 60 artworks. Inter-rater reliability scores were acceptable for each task and experimental condition, but were higher for judges that received instructions to discount technical execution. These and other results are discussed, and the argument presented that, for future CAT studies in this domain, specific instructions to discount technical execution offers a more reliable measure of graphic design creativity.

## ARTICLE HISTORY

Received 28 February 2015

Accepted 6 August 2015

## KEYWORDS

Consensual Assessment Technique; graphic design; design creativity assessment

## 1. Introduction

Within the creative industries, graphic design links across many sectors: be it the need for marketing material; a new logo for an organization; the presentation of scientific information; or the development of a new product. Graphic design will play a part: sometimes in the background, at other times centre stage.

Creativity, in this context, is an asset, and gives clients, design agencies, and individual graphic designers an edge in a competitive market. Whilst not all design opportunities give scope for creativity, many do, and finding novel ways to visually communicate with an audience is often an implicit expectation for competency.

Few would argue assessing graphic design creativity is anything other than a highly subjective process; however, such subjectivity need not be problematic if assessors concur in their subjective judgments. This is exactly what the Consensual Assessment Technique (CAT) maintains to achieve: a research method that can reliably assess creativity, through the consensual assessments of domain experts.

The CAT method is based upon an operational definition of creativity in which "... a product or response is creative to the extent that appropriate observers independently agree it is creative.

**CONTACT** Karl K. Jeffries [kjeffries@uclan.ac.uk](mailto:kjeffries@uclan.ac.uk)

© 2015 The Author(s). Published by Informa UK Limited, trading as Taylor & Francis Group.

This is an Open Access article distributed under the terms of the Creative Commons Attribution-NonCommercial-NoDerivatives License (<http://creativecommons.org/licenses/by-nc-nd/4.0/>), which permits non-commercial re-use, distribution, and reproduction in any medium, provided the original work is properly cited, and is not altered, transformed, or built upon in any way.

Appropriate observers are those familiar with the domain in which the product was created or the response articulated” (Amabile, 1982, p. 1001). The past three decades of CAT research studies have shown, for the most part, that satisfactory levels of agreement are possible across a wide range of domains. Precisely how researchers interpret the CAT method and protocol forms a substantial part of this current study, and are discussed in Section 3.

Despite extensive use within creativity research, the use of the CAT as a measure of creativity within design research is relatively small. Over a 30-year period, for example, only 11 papers were related to design journals (Jeffries, 2012a), and on further review, only two papers operationalized the CAT in their research (Christiaans & Venselaar, 2005; Pektas, 2010). Building upon this 30-year CAT database, follow-up searches, identified 24 papers that made reference to both graphic design and the CAT. On inspection, the vast majority of studies, however, were not related to graphic design directly, with two exceptions (Dineen & Niu, 2008; Silvia et al., 2008).

Silvia et al. (2008) undertook a study to validate a new method of scoring divergent thinking tasks. Part of the study had participants who majored in arts subjects (accounting for 9% of participants in the study), of which some participants majored in graphic design. Whilst the new technique built the case for the validity of subjective rating by citing CAT studies, the method used was not the CAT protocol. Indeed, they acknowledged the importance of expert domain judges for “studies of real creative product” (p.70), but argued that this need not be the case for divergent thinking tasks assessment.

In contrast, Dineen and Niu’s work (2008) utilized the CAT method, and involved participants in their second year of graphic design at a Chinese art and design higher education institution. The study explored the respective merits of UK creative pedagogy relative to traditional Chinese pedagogy. However, it is arguable, how closely aligned to graphic design the final outputs were relative to illustration design. In no way do distinctions that can be made between graphic design and illustration design undermine the value of Dineen and Niu’s research, but for the purposes of this review it highlights that at the time of this study, there was at best only one published paper that had applied the CAT to graphic designers.

The CAT has continued to grow in popularity within design research, and a recent PhD study has applied the CAT to traditional graphic design (Wojtczuk, 2014). A number of findings on the influence of judges’ background on design creativity assessment are highlighted by Wojtczuk, and of particular importance was the low level of consensus achieved by designers’ rating of graphic design creativity.

Therefore, as yet, there is little precedent established for a task to reliably measure graphic design creativity using the CAT. This is particularly relevant in the light of debates surrounding the domain specificity/generalizability of creativity, and the role task selection plays in creativity assessment (Byrne, 2011). These issues raise both practical and theoretical implications for creativity research within and beyond graphic design. Thus, for the first time, this study aims to test two tasks to measure graphic design creativity, and by extension, the reliability of the CAT as a measure of graphic design creativity.

## 2. Pilot studies

Prior to the main study in this paper, two studies were undertaken to explore suitable tasks for a graphic design CAT (Jeffries, 2012b). Each task had current usage in design education, and gave participants the opportunity for graphic design creativity. One used text only, and required participants to choose a word, and then visually communicate that word through the use of type (Figure 1). The other task was to design a graphic which could be transferred to a plain white T-shirt. The graphical image was to be based on the theme of “hands”. For brevity, only a selection of pilot findings will be discussed, particularly those that informed the current research design, but full details of these studies are available on request.

A key finding was that inter-rater reliability was below acceptable levels, or marginal for both tasks: .56 for the T-shirt Task, and even .69 for the Type Task was not ideal. Was something happening in graphic design that warranted caveats for research design, or was this domain challenging the assumptions upon which the CAT was based (i.e. domain experts can independently agree on creativity to an

acceptable level of inter-rater agreement)? The outcome from these pilot studies was that a number of caveats needed to be considered in order to optimize the CAT method as a measure of graphic design creativity: specifically, the influence of technical execution on ratings of graphic design creativity; the background of judges; the range of artwork available; and the suitability of a task for research purposes.

### **2.1. Technical execution**

Artwork in graphic design can vary in its level of technical refinement. At one end of the spectrum are conceptual artworks, where the seed of an idea can be perceived: even if the artworks lack refinement in, for example, font selection, layout, or composition, the creativity of the idea can still be evaluated. At the other end of a technical spectrum is finished artwork; artwork that is ready to go to print or publication, where every aspect of visual communication has been crafted and refined to perfection by the designer.

The pilot studies highlighted that judges' preferences for technical execution differed with regard to judging graphic design creativity. For example, their preference for technical quality appeared to be heightened or subdued relative to the medium of the artwork (hand-drawn sketches, artwork created with the use of computer software, or a combination of both mediums).

Within the CAT literature, some researchers have created instructions that directly address this issue of discounting technical execution (Baer, 1993). Other researchers (the majority of CAT studies) have not done so with little adverse impact: some ask judges to rate creativity alongside technical execution and aesthetic appeal (Amabile, 1982; Christiaans & Venselaar, 2005; Valgeirsdottir, Onarheim, & Gabrielsen, 2015), some only do this the first time they undertake a new CAT task (Hennessey, 1994; Kaufman, Plucker, & Baer, 2008); some do not. Thus, the issue of technical execution ranges from explicit, through to implied, implied only once, or not mentioned at all.

Given this range of approaches, if graphic design experts are not specifically guided to discount technical execution then ambiguous instructions may impact on the level of consensus: could differences in technical preference explain low inter-rater agreement in previous studies? In this study, two sets of instructions will be given to judges to test whether discounting technical execution, or not, has an impact on levels of consensus.

### **2.2. Suitable judges**

In the pilot studies, novice/intermediate judges, relative to experts judges, achieved higher levels of consensus for the Type Task but the rankings, and rationale for ratings, were not the same. The debate over the use of novice or intermediate judges is a contentious one, and likely to remain as such. For future CAT studies in graphic design, the implication was to use domain experts only. A further point is that domain experts can be based within professional practice (i.e. full-time graphic designer) or educational practice (i.e. full-time graphic design lecturers teaching on undergraduate and postgraduate graphic design courses). Previous research in other design domains (Jeffries, 2011) suggest that the values shared on creativity between academics and practitioners is not as polarized as populist views can imply. However, whether this is the case for graphic design is unclear, and thus a cautious approach to CAT studies, in this domain, would be to gather experts from either academia or professional practice, but not to use both within the same group (that is, until research findings can show otherwise). In the present study, only full-time professional graphic designer were used as judges, and each judge was required to have over two years of professional experience within graphic design.

### **2.3. Range of artwork**

Whilst the CAT is a method that requires judges' ratings are relative to other works within a sample, it is feasible that artworks too similar in creative quality pose a more challenging assessment for judges. This lack of diversity may explain the low levels of inter-rater reliability in previous studies. For example,

the artworks gathered for the Type Task were stratified across highly accomplished graphic designers (as this artwork was gathered from the webpages of various designers; few are likely to promote their least creative work).

In contrast, if inter-rater reliability could be low due to sampling from mostly highly creative artworks, then why not for other skewed samples, be they predominantly low or medium in quality? The sampling of T-shirt artworks in the pilot study, for example, were sampled from a module with an introductory nature, thus participants were mostly novices to design, and the resulting artwork may have been clustered towards the lower end of quality. The key issue here is not the quality of individual artwork in isolation, but the range of quality within a sample, and how researchers may inadvertently cluster to a particular range (be it mostly high, or mostly medium, or mostly low quality) relative to the expectations of domain experts.

Given these observations and the fact that to date CAT studies in graphic design had not achieved a suitable level of inter-rater agreement, it was considered prudent that some form of pre-test for diversity of artwork be used in further studies. This would offer a degree of control and insight into how diverse the artwork presented to judges would be, and create optimal conditions under which CAT assessment would take place: if a suitable level of consensus could not be achieved under these experimental conditions then, indeed, something deeply problematic was likely to be occurring in relation to the CAT and graphic design creativity.

#### **2.4. Task selection**

As a final consideration, the T-shirt Task may have been too complicated for the purposes of this study. Firstly, as a task it was designed for an educational purpose with a broad scope, and one that took place over an extended period of time (relative to the more experimental tasks used by many CAT studies). Secondly, the artwork ranged from hand-drawn sketches to artwork created with the use of computer software, such as Adobe Illustrator, InDesign, or Photoshop. Moreover, and thirdly, some artworks incorporated text within the T-shirt graphic, whilst other did not. Each of these issues may have contributed to the low level of consensus achieved in the pilot study. Given this, the implication was for artworks to contain image only, or text only. Furthermore, the medium through which the artwork was created should remain consistent across a sample, and within itself. For example, either all artworks would be hand-drawn, or alternately, created with the use of computer software; a combination of both mediums was to be avoided. Additionally, the Image Task should be able to be completed in a time scale comparable to the Type Task.

Based upon such caveats, the main study was specifically designed to ask: can professional graphic designers achieve inter-rater reliability at or above .7 for an image based graphic design task and a text based graphic design task? If they can, do CAT instructions to discount technical execution increase inter-rater reliability when compared to instructions that make no such stipulation?

### **3. Method**

As is often the custom for CAT studies, two broad groups of participants were required: those who generate the creative outputs (participants) and those who assess the creative outputs (raters, or judges). For many CAT studies outside design, participants are recruited and undertake the creation of a piece of work under experimental conditions – this is because the purpose of such studies is to test the influence of teaching or environmental factors that may impact upon creativity. In this study, it is the judges, rather than the participants that are the focus. Thus, this study follows a research direction set by other researchers (e.g. Baer, Kaufman, & Gentile, 2004; Christiaans & Venselaar, 2005), for the use of work created under non-experimental conditions. It has been showed that judges' inter-rater reliability can remain acceptable even when creative outputs are not generated under experimental conditions.

As a result, graphic design artwork was the creative output, and created as a natural result of engagement with a university degree. Specifically, the study gained consent from participants to use type only and image only artwork created during two assignments for a BA (Hons) Graphic Design

course. These two tasks provided 30 type and 30 image examples to be independently assessed by 16 professional graphic designers using the CAT. The dependent variable was instructions given to judges, and each judge was randomly assigned to receive different instructions for each task. Data was analyzed for inter-rater reliability, and appropriate statistical analysis was used to compare the influence of different instructions on judges' ratings.

### **3.1. Ethical considerations**

As this project was between two universities, authorization was sought, and obtained, from Ethics Committees at each.

One of the core pedagogies of graphic design education is the use of a critical review (commonly known as a "Crit"). During a Crit, students present their finished artworks and have these commented upon by staff and students. Frequently, such Crits make comparisons to other works under review, and opinions of merits and weaknesses are discussed. It is the work under review rather than the designer, but the results of this study could risk reinforcing self-labeling and issues of social standing within a peer group. This was mitigated through a number of measures:

- The result of the study would not use illustrations of any artwork provided. Thus, it would not be possible for the students taking part to know how judges rated their artwork from the results of the study.
- The student artwork collected received an anonymous code at the start of the research. Only the principle investigator knew which codes related to which students, and the principle investigator had no involvement with the assessment of graphic design students during their course of study.

### **3.2. Tasks**

The Type Task was the same as the pilot study. A new Image Task was used that required participants to select two images, and when seen side by side made some sort of creative juxtaposition, or interesting visual communication about the images chosen. As with the Type Task, this has a pedigree within graphic design education and the juxtaposition of images is also a technique frequently used by professional graphic designers.

Whilst none of the artworks created for this study are presented here (as justified above), examples of six artworks previous and post this study are shown below in Figure 1, to give a visual context for these tasks. Specifically, the first Type Task example took the word "Coffee" and through the choice of typeface has rotated the letter "C" to resemble a coffee pot; the second Type Task example, based on the word "Imagine" has deleted the middle letter, leaving the viewer to imagine what this deletion may be; the third type example, has taken the word "Saw", and through extending the last letter "W" implies the teeth of a saw; the first Image Task example placed a pipe wrench next to an image of Robocop; the second a spiral shell next to a satellite image of a tornado; the third, the details of a leaf next to a section of road map.

### **3.3. Instruction to judges**

For each task, whether type or image, two sets of instructions were developed. One set was an adapted version of Kaufman, Baer, Cole, and Sexton's (2008) study, and is cited as an exemplar of CAT instructions (Kaufman et al., 2008).

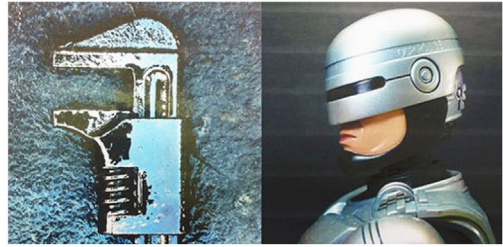
Please look through these artworks, and rate them for creativity. There is no need to explain or defend your ratings in any way; we ask only that you use your own sense of which is more or less creative (relative to the other artworks provided).

Please look through these artworks three times, and rate them for creativity.

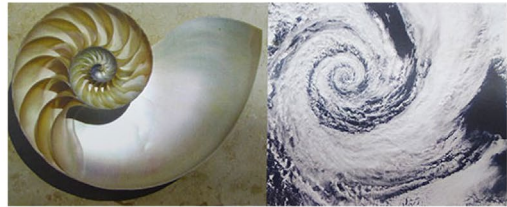
The first time familiarize yourself with all the artworks provided.

The second time, group the artworks into Low, Medium, or High ratings.

offee



ima ine



SAWWWW



**Figure 1.** Six examples of artwork (none of these examples were part of this study): three on the left relate to the Type Task; three on the right relate to the Image Task.

The third time, assign a numerical rating between 1 and 6 (1's being the least creative and 6's being the most creative).

There should be a roughly even number of artworks at each of the six levels. It is very important that you use the full 1-6 scale.

The other set of instructions were exactly the same as Kaufman et al.'s, with one difference, the first sentence (*Please look through these artworks, and rate them for creativity*) was replaced with two sentences adapted from Baer's (1993) CAT instruction for judges (a study where judges were specifically requested not to consider other factors that may impact on storytelling, for example, aesthetic appeal, or grammar, as part of their criteria for rating creativity). Baer's (1993) study is also cited as an exemplar of CAT instruction (Kaufman et al., 2008). The wording used was as follows:

There is only one criterion in rating these artworks: creativity. We realize that creativity probably overlaps other criteria one might consider (for example: aesthetic appeal, or technical execution) but we ask you to rate the artworks solely on the basis of their creativity.

The rationale for the use of Kaufman et al.'s instruction was to provide a clear procedure for judges to follow, and to use instructions that were rating focused rather than ranking focused, as Baer's instructions cited above are (though this does change in other studies related to his 1993 work).

Subsequent additions to Kaufman et al.'s wording were created to clarify the issue of standards; a topic that arose during post-assessment debriefing with graphic designers (and a topic present in the pilot studies). Relative to other artworks, the instruction to rate the "most creative" seemed to be difficult for some judges. Setting aside issues of whether, in the pilot studies, the sample of artwork reflected enough creative diversity, the likelihood for such difficulty was the internal standards a judge had for creativity, and the congruency or discomfort experienced when they were tasked to label

artworks “high” or “most” creative when, relative to the general standards of the domain, they were weak. Clearly, CAT instructions ask for comparison within a group of artwork (as does the theoretical assumptions which underpin this assessment technique) but it could still be the case that for graphic design professionals, mindful of a domain that judges them and others against expectations of high creativity, this is difficult to “turn off”. In this respect, it was felt important to emphasize the relative nature of the ratings, and thus the phrase “... (relative to the other artworks provided)” was added to supplement Kaufman et al.’s instructions.

In contrast, the key distinction to highlight for Baer’s instructions is the emphasis his wording places on one criterion: creativity; and explicit instructions that other factors related to creativity, such as technical execution and aesthetic appeal, are to be discounted. This is a feature absent from Kaufman et al.’s instructions.

Moreover, with Baer’s wording, he instructs judges to use their own “mysterious expert sense” (Kaufman et al., 2008, p. 65) of what creativity is, but to isolate this from other related factors (which he lists). In this way, Baer maintains only one unambiguous criteria is present. This instruction could, however, be viewed as something of a contradiction; does such an instruction not imply other criteria are present and at work: to the extent that they need to be separated from creativity? Like general standards, as mentioned above, the ability to “turn off”, to isolate creativity from other criteria, may be more challenging for some judges and straightforward for others. That said, difficult need not infer that judges are unable to undertake such assessment: what Baer’s original instructions acknowledge is the possibility that this may be a challenge, but one that is necessary and achievable. In this study, Baer’s instructions were edited to two sentences, and reflected a more formal tone in keeping with Kaufman et al.’s.

It is worth noting that whilst Kaufman et al.’s instructions do not discuss other criteria (such as technical execution or aesthetic appeal) and Baer’s make a considerable point of this, previous researchers using CAT have instructed judges to rate creativity alongside, but separate to, aesthetic appeal and technical execution (Hennessey, 1994). Indeed, in Amabile’s (1982) work, the extent to which creativity may be isolated from such factors was a formative part of her paper, and she concluded that

... although judges were not provided with a definition of creativity ... they consistently and reliably identified a quality in both types of product that was distinct from technical execution. Moreover, for artworks, it was distinct from aesthetic appeal as well (p.1010).

At the same time, Amabile acknowledged that for some domains the distinction between technical execution and aesthetic appeal may be less clear, and that creativity is likely to correlate with these aspects of the work. Even within Amabile’s (1982) studies, she found correlations as high as .77 between creativity and technical goodness; and Hennessey (1994) presented statistically significant correlations as high as .71. Contemporary CAT-based research and studies more directly related to professional design, such as Valgeirsdottir, Onarheim, and Gabrielsen (2015), have also identified high positive correlations, yet other design creativity researchers make no mention of having considered this in their research design.

A suggestion, for some time, has been that when a CAT is developed for a different domain, researchers should ask judges to rate both technical execution and aesthetic appeal (Amabile, 1982), and check to see creativity ratings are distinct from these criteria. Once this has been shown to be so, researchers need only ask for ratings of creativity, and can assume that technical execution is no longer a consideration for this task. It is at this point, as stated above, that Baer’s instructions to discount technical execution contrast markedly with those of Kaufman et al.’s. The question is how much does that contrast matter?

### **3.4. Selection of artwork**

In order to select 30 artworks for each task, grades created as a result of student artwork for a creative thinking module were used: the criterion was creativity, and this was assessed by academic staff independent

of this study. The details of this selection process have been reduced for brevity, but the purpose was threefold: firstly, to identify a diverse range of artworks across all CAT levels 1–6; secondly, to have five artworks represented at each CAT level; thirdly, to have the same participant represented in both tasks: type, and image. This would allow CAT scores for type and image to be aggregated to get a total score for both tasks and individuals. As above, the CAT ratings for this study would require judges to rate artwork as low, medium or high, and then rate these from 1 to 6. To determine which artworks, and participants would be selected, each academic grade was stratified to a CAT rating as follows: marks between 44 and below, and 44–50 where rated as low (CAT level 1 and 2); 50–54, and 55–59 as medium (CAT level 3 and 4); 60–69, and 70 and above as high (CAT level 5 and 6). When each of the artwork options were placed alongside each other, this highlighted some CAT levels had fewer options than others. For example, only five artworks were available at CAT level 1 for the Type Task, and only five artworks were available at level 6 for the Image Task. The inclusion of a participant in the Type Task at CAT level 1, for example, determined their representation within the Image Task, and vice versa. Moreover, when a participant was chosen to represent a specific CAT level (regardless of which task), this influenced the options available for other CAT levels. In this respect, the choices for selection became to an extent self-identifying, with limited options depending on whether a CAT level had more than five artworks available. The process was iterative, and became progressively more challenging with each inclusion. However, the stratification was achieved, and 30 participants were identified whose text and image artworks represented each of the six CAT levels, with five artworks at each level. For reasons of research ethics and participant confidentiality, examples of the artwork used in this study have been withheld from publication.

### **3.5. Judges**

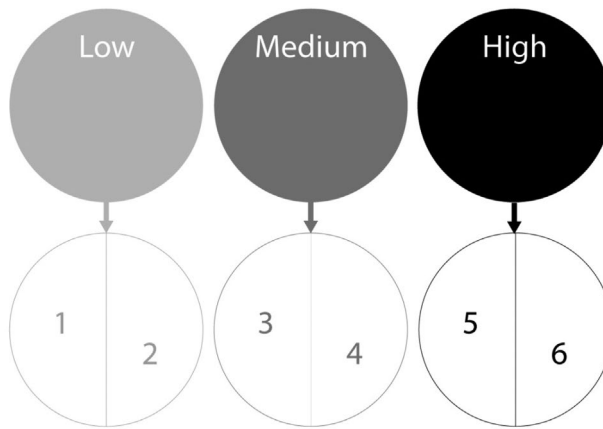
Previous research guidelines suggest that for CAT reliability, between 5 and 10 judges is an acceptable number for a given task. For this study, 16 full-time professional designers took part (sampling detail are available in the results section). Judges were randomly assigned to two groups, in which task and instructions were rotated to counter balance whether they receive the Image Task first or second, with whether they received Kaufman et al.'s or Baer's instructions, or a combination. This was to minimize order effects, particularly practice effects and fatigue effects.

### **3.6. Procedures**

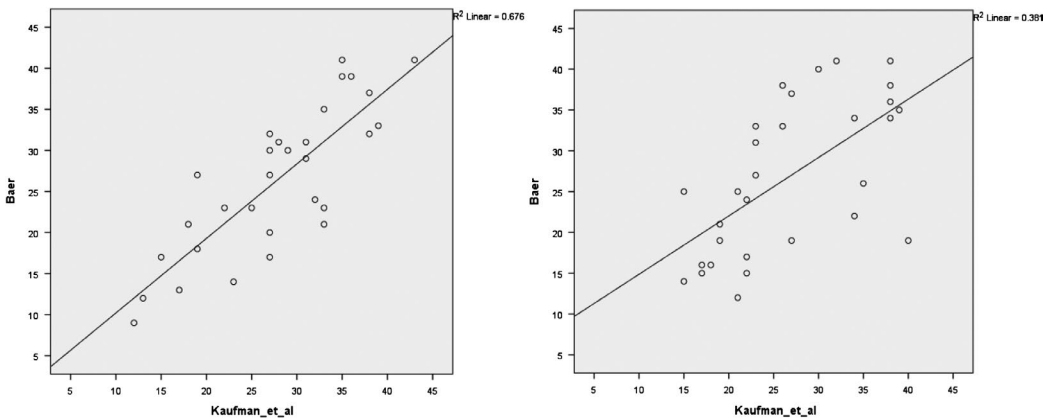
The procedures for rating artwork were the same for each judge and task. Initially, judges answered three questions: their years of experience in graphic design, whether they would describe themselves as a graphic designer, and their age. After this, each judge was given the instructions for their first task, alongside an example of the design brief given to participants for this task. They had as much time as they required to read the instruction. Next, each judge accessed a laptop with a PDF slide presentation of the 30 artwork for the first task. The order of artwork was randomized, and they were free to control how long they viewed artworks, and could return to each artwork for further inspection. Each judge familiarized themselves with all the artworks, and when satisfied informed the researcher they were ready to continue. Judges were given an A3 laminated rating sheet, see Figure 2 (developed to graphically reinforce Kaufman et al.'s CAT protocol and instructions), and a set of laminated cards. These cards were miniature copies of the artwork they had just viewed. Cards were placed in a stack, by the researcher, onto the rating sheet area designated "Medium", and judges proceeded to rate the artwork, and had as much time as they required. Task two followed the same procedures. All judges were debriefed on the purpose of the experiment, and had the opportunity to ask any questions about the study.

## **4. Results**

Within a year group of 66 students, 48 students gave their consent to take part in the project. The median age was 19 years (SD 1.46); 18 female and 30 male students took part. The mean age for judges



**Figure 2.** Rating sheet developed to graphically reinforce Kaufman et al's CAT protocol.

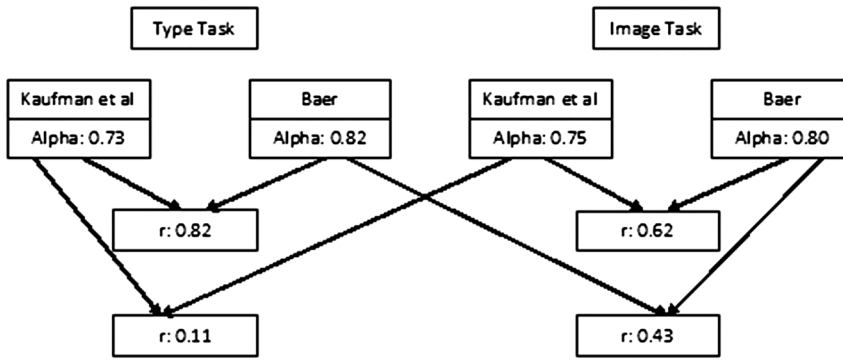


**Figure 3.** Scatterplots for Type Task (left) and Image Task (right).

was just over 41 years of age (SD 9.80), and ranged from 30 to 63 years of age; six judges were female. All judges identified themselves as graphic designers, and their professional experience within graphic design ranged from 7 to 35 years: the mean being just over 17 years.

For the Type Task, the eight judges who received the adapted Kaufman et al. CAT instructions had an alpha of .73; for the eight judges who received the adapted Baer's instruction to discount technical execution the alpha was .82. The skew for aggregated ratings for each group were within acceptable limits at the less than 5% level: adapted Kaufman et al. z-skew was  $-.62$ ; adapted Baer z-skew was  $-.2$ . Pearson's  $r$  was  $.82$ , suggestive of a very strong positive correlation between the scores, and was a significant correlation at the .01 level (two tailed). The scatter plot on the left (Figure 3) shows the regression line and strength of correlation.

For the Image Task, the eight judges who received the adapted Kaufman et al's CAT instructions had an  $\alpha$  of .75; for the eight judges who received the adapted Baer's instruction to discount technical execution the  $\alpha$  was .80. The skew for aggregated ratings for each group were within acceptable limits at the less than 5% level: adapted Kaufman et al. z-skew was  $.74$ ; adapted Baer z-skew was  $.10$ . Pearson's  $r$  was  $.62$ , suggestive of a strong positive correlation between the scores, and was a significant correlation at the .01 level (two tailed). The scatter plot to the right (Figure 2) shows the regression line and strength of correlation.



**Figure 4.** Alpha and  $r$  in relation to task and instructions.

The correlation between the adapted Baer's instructions for both type and image was an  $r$  of .43, which can be considered a moderate positive correlation and one that was significant at the .5 level (two tailed). For the adapted Kaufman et al. instructions for both type and image  $r$  was .11, suggestive of a negligible positive correlation, and one that was not statistically significant. Figure 4 highlights  $\alpha$  and  $r$  in relation to which task and instructions.

## 5. Discussion

Given the novelty of the CAT for the assessment of graphic design creativity, there was a need to pilot whether judges showed a suitable standard of consensus as achieved in other research. Of the few studies directly related to design, the CAT has shown sufficient levels of consensus within their respective domains. It was expected that using the CAT to assess graphic design creativity should follow a similar pattern of inter-rater reliability; however, the pilot result did not achieve this. A number of reasons, such as the range of artwork, technical preferences of judges, task selection, and the sampling of judges were accounted for in the present study, and appear to have resolved the previous issue of inter-rater reliability. In either task, the inter-rater reliability was acceptable, all were above .7, the highest being .82. Given this, these particular tasks can be considered reliable measures of graphic design creativity using the CAT. Prior to this study, the choice of tasks to measure graphic design creativity was not obvious, and the pilot findings highlighted that some task do not translate well from design education context into experimental research.

Whilst each task (regardless of the instruction to judges) had acceptable alphas, a marked difference can be seen in the correlation between the Type Task ( $r = .82$ ) and the Image Task ( $r = .62$ ). The reasons for this difference could be numerous, complex and interrelated. Perhaps, the Type Task, and typography, shares a common knowledge base for graphic designers; possibly the Type Task is less influenced by discounting technical execution than the Image Task; the inclusion of color in the Image Task may add to the complexity of assessing its creativity: color was absent in the Type Task. What can be said, and has been said by other researchers (Reiter-Palmon et al., 2009), is that task selection is an important factor in creativity assessment, and our depth of understanding is "essentially missing in the literature" (Lubart & Guigard, 2004, p. 48).

In early CAT research, Amabile (1982, 1996) concluded that judges were able to distinguish creativity from other aspects such as aesthetic appeal and technical execution. Does this finding still apply to graphic design creativity? Whilst the differences between the adapted Kaufman et al. instructions and the adapted Baer instructions (to discount technical execution) ranged from .05 to .09, the difference was towards higher levels of inter-rater reliability when judges were asked to discount technical execution from their creativity ratings; this occurred in both the Type Task and Image Task. Arguably, these differences are slight, but it may be that acceptable inter-rater reliability is not enough in isolation, and that the other consideration is the correlation between aggregated scores. Only the adapted Baer

instructions were statistically significant, and suggestive of a moderate to strong positive relationship. A further finding was that correlations were significant at the .01 level in either task, but were stronger for the Type Task, than the Image Task. It appears the Type Task is less influenced by discounting technical execution, and the Image Task may be more susceptible.

Whilst the CAT is a method that requires ratings are relative to other works within a sample, it is feasible that artworks too similar in quality pose a more challenging task for judges than those that show more diversity. Such a possibility is interesting. Several CAT studies highlight the real world basis of the technique, and with acceptable inter-rater reliability, there is little reason to question its reliability. However, most CAT studies are research studies, and whilst parallels can be drawn between CAT methods and those used by judges of, for example, professional competitions, awards, or traditional art school degree show assessment, the independence of judges does not happen throughout the rating process as it does in the CAT.

Much like other consensual techniques, such as the Delphi technique, competition judges do tend to confer with each other towards the end of the process. A judge's initial assessment may be independent of the panel, but towards the final stages of evaluation, debate and compromise is not uncommon in order for consensus to be achieved. This need not discredit CAT methodology, it does suggest, however, that the real world assessment of highly creative artworks may require more debate amongst judges than the CAT currently allows.

If this is the case for highly creative artworks, then why not for other skewed samples be they predominantly low or medium in quality? It is beyond the scope of this research to explore this in detail, but the argument, given the pilot findings, and the subsequent main study, is that diversity of artwork may play a significant factor in levels of consensus amongst judges. Therefore, it is possible for researchers to assume enough diversity exists in a sample of artwork, when it may not.

Precisely how researchers interpret the CAT method and protocol is open to debate (Amabile, 1982; Baer, 1993; Hennessey, 1994; Kaufman et al., 2008). In essence, the argument presented here is to develop standard CAT instructions to rate creativity for all design creativity researchers: those who choose to measure additional criteria, such as technical execution and aesthetic appeal, in each study; those who do so only once; those that measure only creativity. In design domains where the distinction between creativity, technical execution and aesthetic appeal appears to be a rather thin one, the findings of this paper suggested that explicit CAT instructions to discount technical execution from creativity ratings are more reliable.

The main point to consider is whether to include a caveat around technical execution in future research. By inclusion, such a caveat, directly addresses assumptions around technical execution and creativity. Indeed, if the CAT is foremost a measure of creativity (however judges interpret this word), then clarification on technical execution seems a reasonable distinction to bring to their attention. The slight increase in inter-rater agreement for judges that received the technical execution caveat can be interpreted both for and against its inclusion. More revealing is the correlations between type and image scores relative to instructions. It is only scores where a caveat was included that enabled an aggregated graphic design creativity score. As the purpose of these tasks was to evaluate graphic design creativity by isolating two distinct features of graphic design (the creative use of type and the creative use of image), the expectation was that a degree of positive correlation would be likely in the combination of these tasks. What is interesting to note for this study is that when judges assess exactly the same tasks and exactly the same artwork, only the caveat on technical execution offers the opportunity for an aggregated score, and thus, at least for research purposes, enables distinctions within a group on levels of graphic design creativity.

## 6. Conclusion

Prior to this study, the choice of tasks to measure graphic design creativity was not obvious. In this study, a number of research design factors, such as diversity of artwork, technical preference, task selection, and sampling of judges were accounted for in the research design, and appear to have

resolved previous issues of inter-rater reliability. However, the arguments presented in this paper suggest instructions to discount technical execution from judges' creativity assessment do appear to influence the reliability of the CAT. The difference was towards higher levels of inter-rate reliability when judges were asked to discount technical execution from their creativity ratings; this occurred in both the Type Task and Image Task used. Moreover, only CAT assessments undertaken where the technical caveat was included enable an aggregated graphic design creativity score for both the image and Type Tasks. Perhaps, these implications apply not only to graphic design, but have relevance for all CAT assessments of design creativity? To paraphrase Nickerson (1999), as researchers we have two choices: include a caveat on technical execution that future research will show was not fundamental, or exclude it and find technical execution does influence rating on design creativity. Unless there is some detrimental effect (which does not appear to be the case in this study) then a cautious approach would be for future CAT usage to include a caveat on technical execution when applied to design creativity research.

## Acknowledgments

I would like to acknowledge Dr Alison Green and Dr Theodore Zamenopoulous for their constant support and encouragement, and Dr Gini Harrison for her specific contribution towards the research design; Mr Andrew Bainbridge for enabling access to UCLan graphic design students, classes, and discussion of creative tasks for graphic design; Prof. Lubaina Himid for supporting my attendance at the 3rd ICDC, and so much more; the ICDC and IJDCI reviewers for their insightful comments. Lastly, most of my thanks must go to all the participants (both professional graphic designers and students of graphic design); for reasons of anonymity you each remain unknown here, but you know who you are, as I do, and you have my heartfelt thanks; without your support none of this research would be possible.

## Disclosure statement

No potential conflict of interest was reported by the author.

## Funding

This research was supported by the University of Central Lancashire.

## ORCID

Karl K. Jeffries  <http://orcid.org/0000-0002-8936-1800>

## References

- Amabile, T. M. (1982). Social psychology of creativity: A consensual assessment technique. *Journal of Personality and Social Psychology*, 43, 997–1013.
- Amabile, T. M. (1996). *Creativity in context: Update to the social psychology of creativity*. Boulder, CO: Westview Press.
- Baer, J. (1993). *Creativity and divergent thinking: A task-specific approach*. Hillsdale, NJ: Lawrence Erlbaum Associates.
- Baer, J., Kaufman, J. C., & Gentile, C. A. (2004). Extension of the consensual assessment technique to nonparallel creative products. *Creativity Research Journal*, 16, 113–117.
- Byrne, C. (2011). *Task selection and the consensual assessment technique: Using collage tasks in creativity research* (Unpublished masters thesis). University of Central Lancashire, Preston.
- Christiaans, H., & Venselaar, K. (2005). Creativity in design engineering and the role of knowledge: Modelling the expert. *International Journal of Technology and Design Education*, 15, 217–236.
- Dineen, R., & Niu, W. (2008). The effectiveness of western creative teaching methods in China: An action research project. *Psychology of Aesthetics, Creativity, and the Arts*, 2, 42–52.
- Hennessey, B. A. (1994). The consensual assessment technique: An examination of the relationship between ratings of product and process creativity. *Creativity Research Journal*, 7, 193–208.
- Jeffries, K. K. (2011). Skills for creativity in games design. *Design Studies*, 32, 60–85.
- Jeffries, K. K. (2012a). Amabile's Consensual Assessment Technique: Why has it not been used more in design creativity research? *Proceedings of the 2nd International Conference on Design Creativity (ICDC2012)*, 1, 211–220.

- Jeffries, K. K. (2012b). *Skills for creativity in graphic design* (Unpublished thesis). The Open University, Milton Keynes, UK.
- Kaufman, J. C., Baer, J., Cole, J. C., & Sexton, J. D. (2008). A comparison of expert and nonexpert raters using the consensual assessment technique. *Creativity Research Journal*, 20, 171–178.
- Kaufman, J. C., Plucker, J. A., & Baer, J. (2008). *Essentials of creativity assessment*. Hoboken, NJ: John Wiley & Sons.
- Lubart, T., & Guignard, J. H. (2004). The generality-specificity of creativity: A multivariate approach. In R. J. Sternberg, E. L. Grigorenko & J. L. Singer (Eds), *Creativity: From potential to realization* (pp. 43–56). Washington, DC: American Psychological Association.
- Nickerson, R. S. (1999). Enhancing creativity. In R. J. Sternberg (Ed.), *Handbook of creativity* (pp. 392–430). Cambridge: Cambridge University Press.
- Pektas, S. T. (2010). Effects of cognitive styles on 2D drafting and design performance in digital media. *International Journal of Technology and Design Education*, 20, 63–76.
- Reiter-Palmon, R., Illies, M. Y., Cross, L. K., Buboltz, C. B., & Nimps, T. (2009). Creativity and domain specificity: The effect of task type on multiple indexes of creative problem-solving. *Psychology of Aesthetics, Creativity, and the Arts*, 3, 73–80.
- Silvia, P. J., Winterstein, B. P., Willse, J. T., Barona, C. M., Cram, J. T., Hess, K. I., ... Richard, C. A. (2008). Assessing creativity with divergent thinking tasks: Exploring the reliability and validity of new subjective scoring methods. *Psychology of Aesthetics, Creativity, and the Arts*, 2, 68–85.
- Valgeirsdottir, D., Onarheim, B., & Gabrielsen, G. (2015). Product creativity assessment of innovations: Considering the creative process. *International Journal of Design Creativity and Innovation*, 3, 95–106.
- Wojtczuk, A. (2014). *Creative product assessment in design: Influences of judges' backgrounds and levels of experience in design* (Unpublished doctoral dissertation), Aix Marseille Université, Marseille.